Performance Matrix.

R squared

$$= 1 - \frac{SS_{Res}}{SS_{Tatal}}$$

SS_{Res} : Sum of sq. residuals

SS_{Tatal} : Sum of sq. Average.

$$\boxed{= 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}}$$

y_i : Actual points

$\hat{y}_i$: Predicted points

$\bar{y}_i$: Average of points.

In general $\bar{y}_i > \hat{y}_i$ because of which denominator will be heigher.

Thus, we have :

$$R\text{-squared} = 1 - \frac{\text{Small value}}{\text{large value}}$$

$$= 1 - \text{Small value.}$$

Thus, value of R-squared ≤ 1

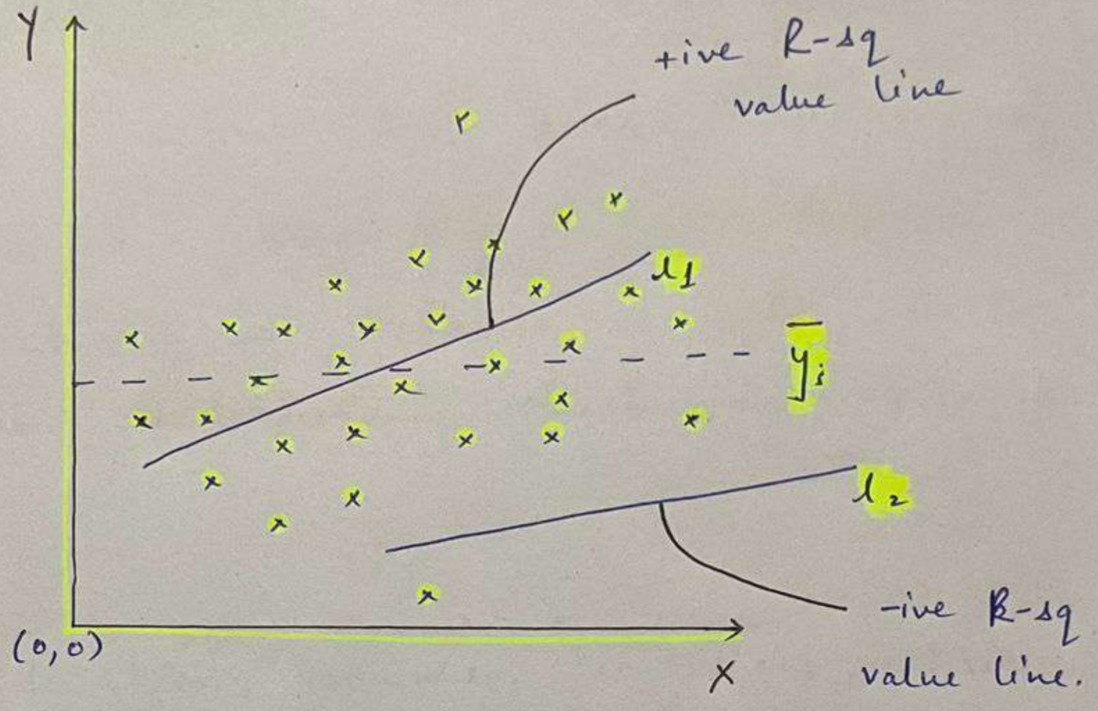
If it comes like .85, .63, .90 means model is 85%, 63% accurate.

Note:

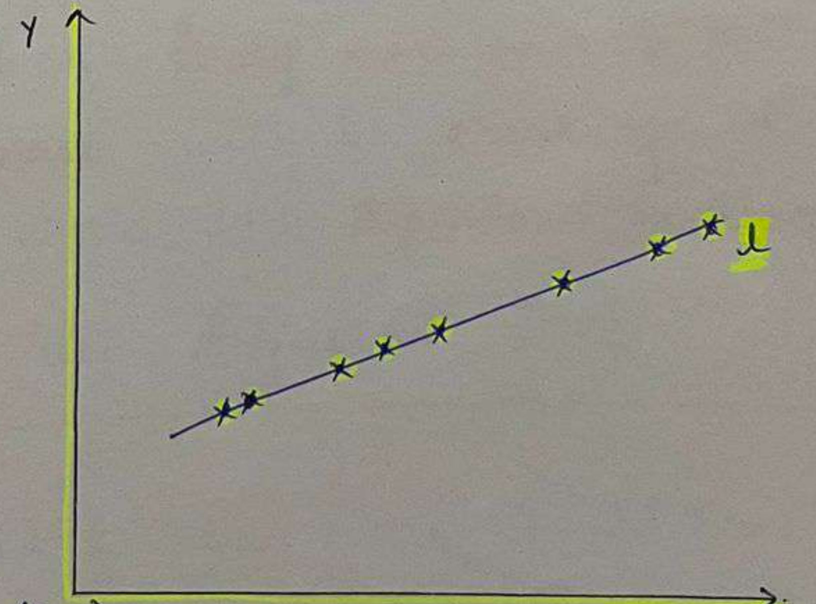
Denominator will be large bcz we are considering average of $\bar{y}_i$ and in case most points will not be around/close to the best fit line bcz of which error will be larger.

- R-square is used to test performance of model being trained/created.

In case if denominator is less then the value of R-square is negative Mean while model is very bad.



Above diagrams shown the condition where $(y_i - \hat{y}_i)$ can be greater as the best fit line is far away predicted point and actual one difference will be larger.



Above diagram shows condition when R-sq = 1 when all data point lies on the line.

Adjusted R-sq.

we use adjusted - R-sq because in case of R-square the accuracy increase/decrease withe change in column of data which is even independent which must not vary.

Ex -

Area, location, Room no.s, People living inside, Price.

consider above data case to decide the price of house. In above How many people living in room does nothing in deciding room price.

In case of R-sq more no.s of feature we select more will be accuracy, it doesn't matter eithe it's dependent or independent.

So, we use Adjusted - R-sq:

$$\boxed{\text{Adjusted } R^2 = \frac{1-(1-R^2)(N-1)}{N-P-1}}$$

N: No.s of dat points | P: No. of Independent feature.

Overfitting & Underfitting

Consider we have 1000 datapoints which will further be separated for training and Testing.

Sample [1000 points]

[750 points Training data]

[250 Points Test data.]

[600 points Train the model]

[150 points validation of the model.]

Hyperparameter tunning of the model.

Assume from the dataset we created the model with having following Accuracy level :

Data Type	Accuracy	
Training data	95% [Excellent]	→ low Bias
Test data	89% [V. Good]	→ low Variance

In case above case occur model will be good as there is not much d/f b/w Training and Test accuracy.

Case I :

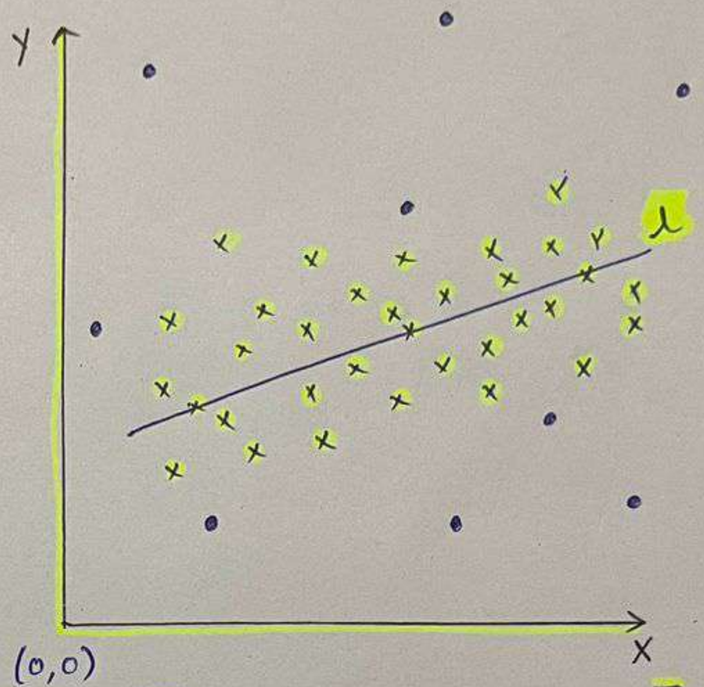
for bellow case.

Data Type	Accuracy.	
Training Data	90% [V. Good]	→ Low Bias
Test data	48% [Poor]	→ High Variance.

For above like situation we call it as Overfitting.

EX - Studied all syllabus but fail in Exam.

Overfitting:



It's clear from fig 1 that the model will predict well for Train datapoints but poor for Test datapoints.

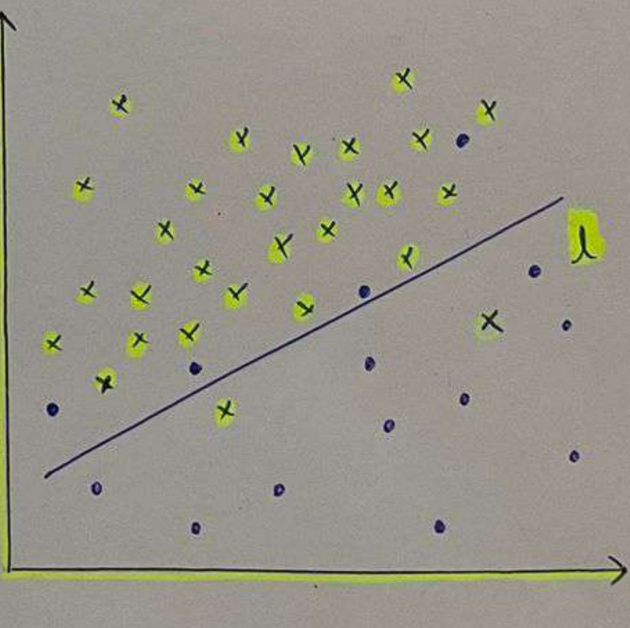
Case II :

for bellow case.

[* - Training datapoints
• - Test datapoints]

Data Type	Accuracy	
Training data	35%	→ High bias
Test data	25% or 85%	→ Low / High variance.

Underfitting :



It's clearn from fig 2 that result for both datapoints is poor but some time might be good.

EX- Random MCQ Ans.